Methods and results from four scientific networking projects

Malathi Veeraraghavan
University of Virginia
mvee@virginia.edu
IEEE ANTS 2013
Dec. 16, 2013

PhD students: Zhenzhen Yan, Jie Li, Zhengyang Liu, Tian Jin
Collaborators: Chris Tracy, ESnet, Steve Emmerson, UCAR, John Dennis, NCAR, Robert D. Rusell, UNH
Web site: http://www.ece.virginia.edu/mv/html-files/research.html

Thanks to the US DOE ASCR for grants DE-SC0007341 & NSF for grants, OCI-1127340 , CNS-1116081, and ACI-1340910

Outline

- Interested in collaborations
- Background
  - National labs & RENs
- Methods & Results: 4 projects
  - Reducing file-transfer throughput variance
  - Traffic engineering elephant flows
  - Reliable multicast of meteorology data
  - Improving execution time of Climate applications (5000 cores)

Graduate Research Assistant (GRA) openings

- GRA work activities include the following:
  - run software programs such as GridFTP/MPI programs
  - write shell scripts, plan and execute experiments on national research testbeds and HPC systems
  - collect measurements, statistical analysis of results
  - literature search and write papers
- Useful skills/languages:
  - Linux, C++, Java, Perl, R, shell scripts, MPI programming
- Useful courses:
  - Software development methods, Computer Networks, Statistics, Parallel Programming
- Email mvee@virginia.edu

3

Scientific community

- US Department of Energy (DOE) funds fundamental research in basic sciences
  - High energy physics
  - Basic energy sciences
  - Biological and environmental research
  - Fusion energy Sciences
  - Nuclear Physics
  - Advanced Scientific Computing Research (ASCR)
- National Science Foundation Office of Cyber Infrastructure (NSF OCI)
  - University Corporation for Atmospheric Research (UCAR)
  - National Center for Atmospheric Research (NCAR)
  - Teragrid and now XSEDE

Labs, universities, Research & Education Networks (REN)

- DOE national labs
  - Oak Ridge National Lab (ORNL), Tennessee
  - Argonne National Lab (ANL), Chicago
  - Lawrence Berkeley National Laboratory (LBNL), Bay Area
  - National Energy Research Scientific Computing Center (NERSC)
  - and other labs …
- Universities – Physicists, Biologists, Earth Sc. et.
- Supercomputing facilities
  - ALCF (ANL), OLCF (ONRL), NERSC
  - TACC (Texas), NWSC (Wyoming), etc.
- ESnet: 100 Gb/s backbone REN for DOE labs
- Internet2: Backone REN for universities/labs

Research and Education Nets

- National Knowledge Network
  - http://www.nkn.in
- Education & Research Network (ERNET)
  - http://www.eis.ernet.in
- China Education and Research Network (CERNET2)
- GEANT2 (European)
- JGN-X (Japan)
- Canarie (Canada)

Communication needs of scientific community

- Large dataset movement across WAN
  - Higher the rate, the better. Relentless!
- Wide-area file system
  - Leave data on local cluster, use remote computational resources
- Remote visualization and computational steering
  - Viz displays at remote site
  - Change parameters to run next set of simulations
- Remote instrument control
  - Beamline control of Photon source
- Intra-datacenter: low-latency comm. for MPI applications - tightly coupled tasks

Outline

- Background
  - National labs & RENs
- ➢ Methods & Results: 4 projects
  - Variance in file-transfer throughput
  - Traffic engineering elephant flows
  - Reliable multicast of meteorology data
  - Improving execution time of Climate applications (5000 cores)

Our group's methods

- Talk to CS programmers/admins who support scientists
  - Identify current problems in scientific computing and networking community
- Submit proposal with the "customer"
- If awarded,
  - customer provides data, access to computer systems
  - reviews analysis results, engineering solution
  - adopts solution (near-term impact)

9

Collaborators/systems used in our projects

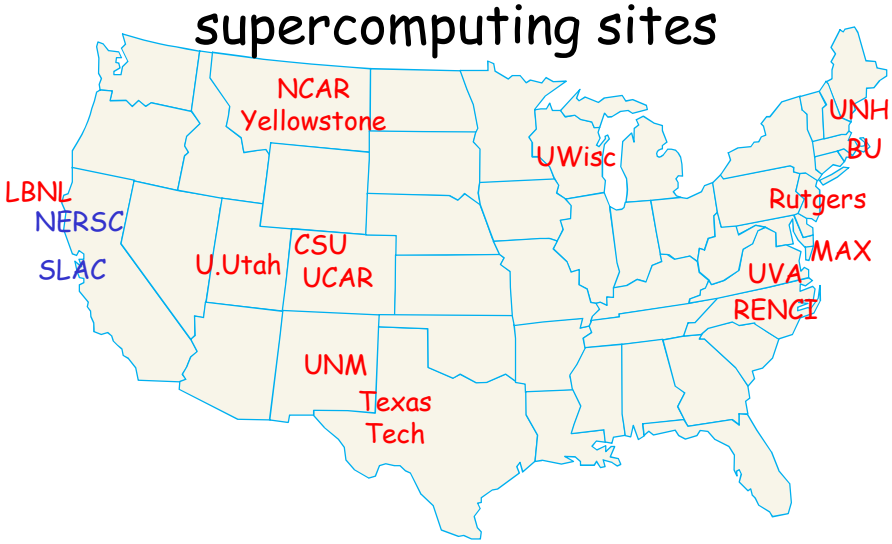
10

Outline

- Background
  - National labs & RENs
- Methods & Results: 4 projects
  - Variance in file-transfer throughput
  - Traffic engineering elephant flows
  - Reliable multicast of meteorology data
  - Improving execution time of Climate applications (5000 cores)

11

NERSC and SLAC: supercomputing sites



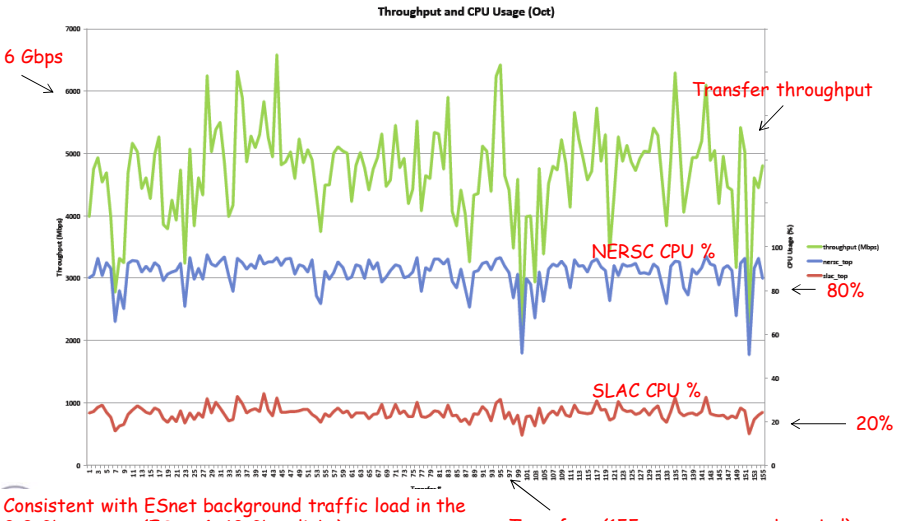
Problem: Find causes of file-transfer throughput variance
Partners: J. Hick (NERSC) and Y-T. Li (SLAC)

12

Instrumented transfers between production DTNs

- Monitoring scripts: initiates *top* (CPU usage) and *tcpdump* (packet loss) before each transfer
- Scheduled hourly GridFTP transfers between NERSC and SLAC DTNs (Data Transfer Nodes)
- Stop data collection after transfers
- Analyze logs collected from GridFTP and monitoring tools
- Method: different from that used in previous papers in which independent sources of measurements were obtained (e.g., NWS for network)

Transfer throughput dependence on CPU time



Regression Model

- Dependent variable: throughput
- Independent variables:
  - NERSC CPU usage
  - packet loss rate
  - SLAC CPU usage?
    - SLAC CPU usage was highly correlated with NERSC CPU usage: linear model

Results

Variable	Mean value
Retransmission Rate (p_i)	4.1E-05
$f(p_i)$ in Mbps	-683.3
Throughput (Gbps)	4.226
NERSC CPU usage (%)	78.28
SLAC CPU usage (%)	21.45

$$y_i = \beta_1' NERSCcpu_i + \beta_2' \epsilon_i + f(p_i) + e_i,$$

Mean values of regression coefficients: $\beta'_1 = 62.708$, $\beta'_2 = 153.194$

Mean value: 62.708 * 78.28 + 153.194 * 0 - 683.3
= 4.9 Gbps – 0.683 Gbps = 4225.48 Mbps

- CPU usage was the primary factor in determining throughput
- However packet loss rate, while small, contributes to throughput reduction (683.3/4225.48 = 16%)

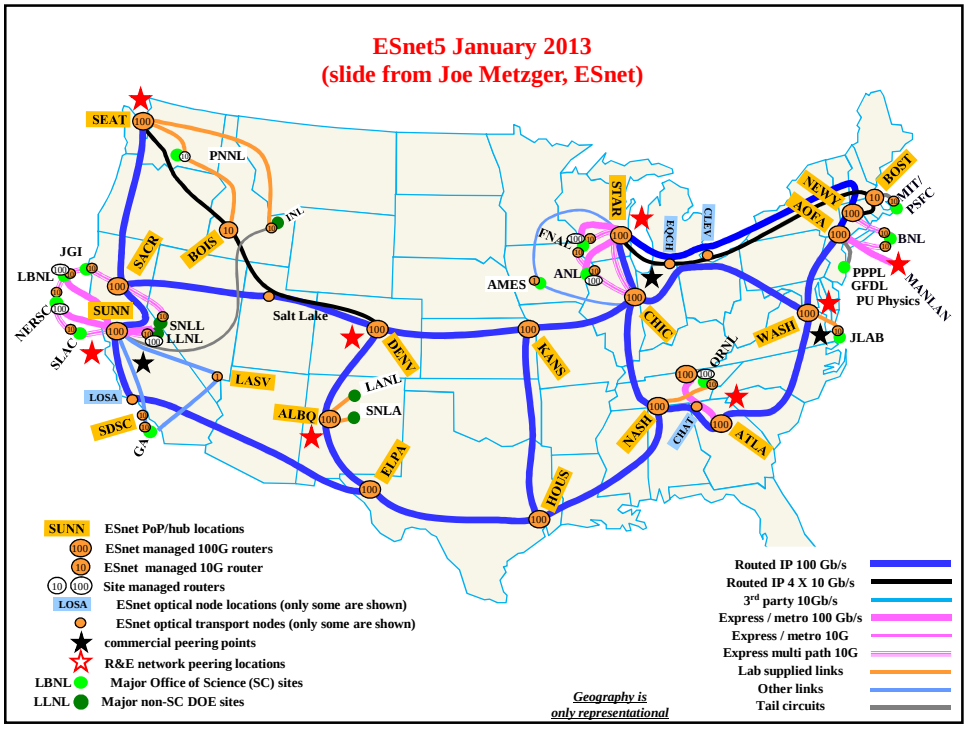
Key finding

- To control variance, the number of concurrent processes on the DTNs have to be controlled via PBS like schedulers
- Current approach:
  - DTNs are used in interactive mode
  - As users login to DTNs and initiate file transfer apps as needed, the amount of CPU and disk resources available to a particular transfer are not controlled

Z. Liu, M. Veeraraghavan, J. Zhou, J. Hick, Y-T. Li, "On causes of GridFTP throughput variance," in Proc. of IEEE/ACM NDM workshop 2013.

Outline

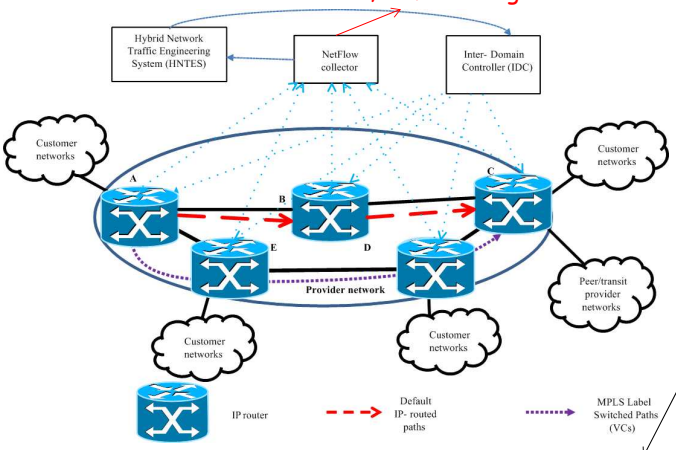
- Background
  - National labs & RENs
- Methods & Results: 4 projects
  - Variance in file-transfer throughput
  - ➢ Traffic engineering elephant flows
  - Reliable multicast of meteorology data
  - Improving execution time of Climate applications (5000 cores)



Hybrid network traffic engineering system (HNTES)

- Intradomain identification/redirection of alpha flows

Located at LBNL, CA; have login access



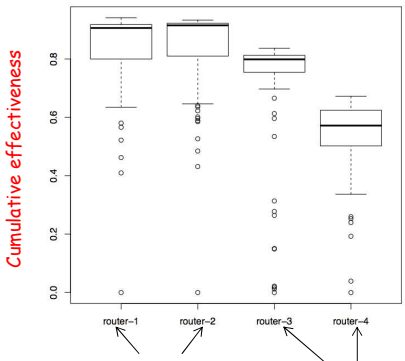
- Three steps
  - Analysis of NetFlow reports from ingress routers to identify address prefixes of completed alpha flows
  - IDC creates L3 circuits between ingress-egress router pairs and configures QoS
  - IDC sets firewall filters to direct future alpha flows with matching address prefixes to L3 circuits

Aging parameter (A): age out rules corresponding to prefixes for which no alpha flows have been observed

20

Cumulative effectiveness (/24)

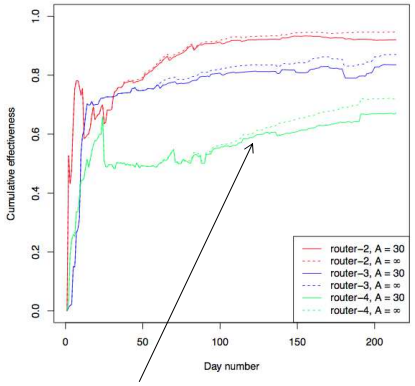
Boxplots for 214 values each



Provider edge routers (single customers)

Peering routers (router-3: REN; router-4: commercial)

router-1 omitted as it is similar to router-2



Why is cumulative effectivness lower for peering routers, esp. router-4?

Observations

- Higher effectiveness for routers 1 & 2:
  - monitored links: incoming side of interfaces from DOE labs
  - downloads from supercomputing facilities are repetitive (a scientist accesses the same data transfer nodes)
- Lower effectiveness for routers 3 & 4:
  - monitored links: incoming side of interfaces from peering networks
  - fewer uploads to DoE labs than downloads from DOE labs
  - expect few, if any, scientific data transfers from commerical peers (router-4)

Key findings

- Offline HNTES solution effective for downloads from DOE labs
- Less effective for uploads esp. from commercial peering links
  - May need online solution
- HNTES extended to characterize size, duration and average rate of alpha flows
- Currently testing HNTES for deployment in ESnet with C. Tracy (ESnet co-PI)

T. Jin, C. Tracy, M. Veeraraghavan, Z.Yan,
"Traffic Engineering of High-Rate Large-Sized Flows,"
IEEE HPSR 2013

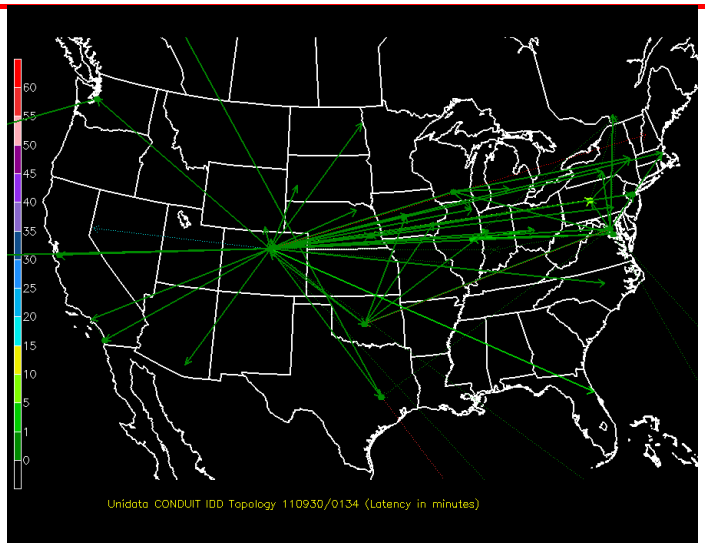
23

Outline

- Background
  - National labs & RENs
- Methods & Results: 4 projects
  - Variance in file-transfer throughput
  - Traffic engineering elephant flows
  - ➢Reliable multicast of meteorology data
  - Improving execution time of Climate applications (5000 cores)

24

CONDUIT data type (example: 163 hosts)
Problem: unicast TCP connections
UCAR receives over 11 GB/h,transmits over 600 GB/h



25

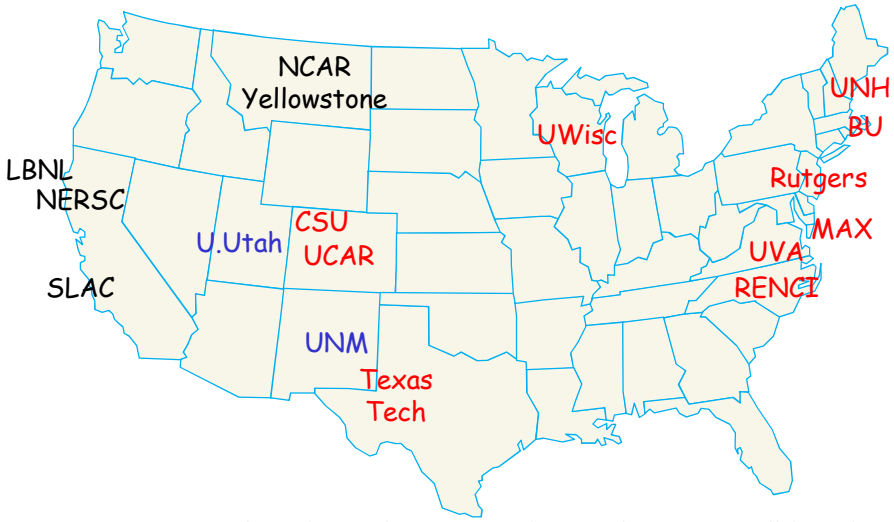
Current status

- Reliable Virtual Circuit Multicast Transport Protocol (VCMTP) designed, prototyped, tested on U. Utah and UNM clusters
  - Tradeoff throughput with robustness
- VCMTP is being integrated by S. Emmerson (UCAR) with LDM software (used for data distribution)
- Will be tested on DYNES (distributed instrument)

J. Li, M. Veeraraghavan, S. Emmerson, R. D. Russell, "VCMTP: A Reliable Message Multicast Transport Protocol for Virtual Circuits," under review IEEE TPDS

26

DYNES sites in red



Partners: J. Robodaiek (UWisc), S. Decker (Rutgers), R. D. Russell (UNH), T. Lehman (MAX); D. Starobinski (BU), Chris Heermann (RENCI), Alan Sill (TT), Greg Redder (CSU), Pete Siemsen (UCAR)

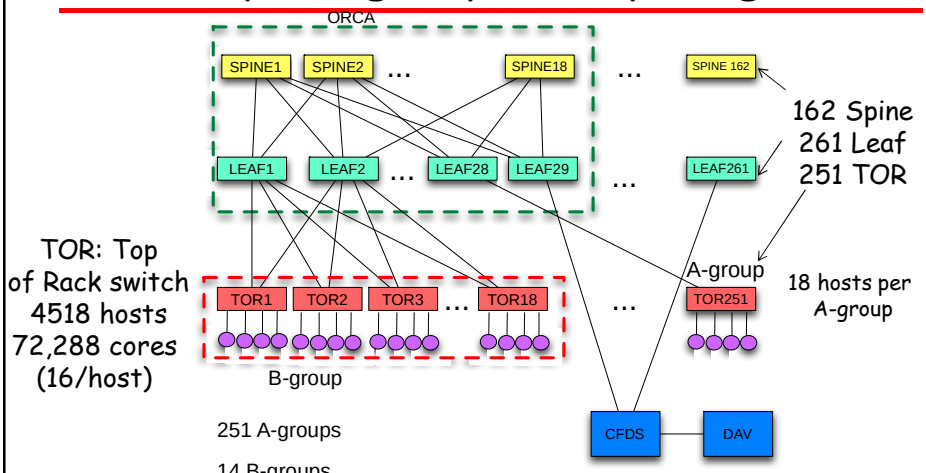
27

Outline

- Background
  - National labs & RENs
- Methods & Results: 4 projects
  - Variance in file-transfer throughput
  - Traffic engineering elephant flows
  - Reliable multicast of meteorology data
  - ➢Improving execution time of Climate applications (5000 cores)

28

Yellowstone: InfiniBand Cluster
NCAR Wyoming Supercomputing Center



Summary

- Brief overview of the high-speed networking needs of the scientific community
- Approach:
  - Collect & Analyze data (collaborators: key)
  - Design solutions to solve a problem
  - Leverage community-wide shared resources (CRI and MRI NSF funding)
- Questions/comments? mvee@virginia.edu